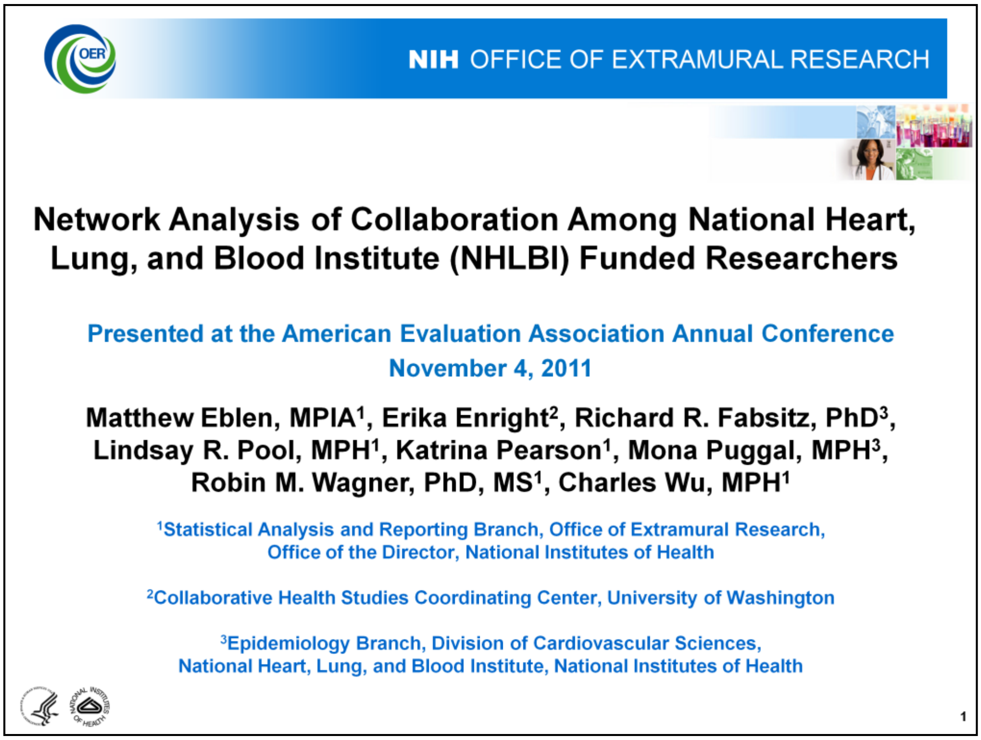


NIH OFFICE OF EXTRAMURAL RESEARCH

Network Analysis of Collaboration Among National Heart, Lung, and Blood Institute (NHLBI) Funded Researchers

Presented at the American Evaluation Association Annual Conference
November 4, 2011

Matthew Eblen, MPIA[1], Erika Enright[2], Richard R. Fabsitz, PhD[3], Lindsay R. Pool, MPH[1], Katrina Pearson[1], Mona Puggal, MPH[3], Robin M. Wagner, PhD, MS[1], Charles Wu, MPH[1]

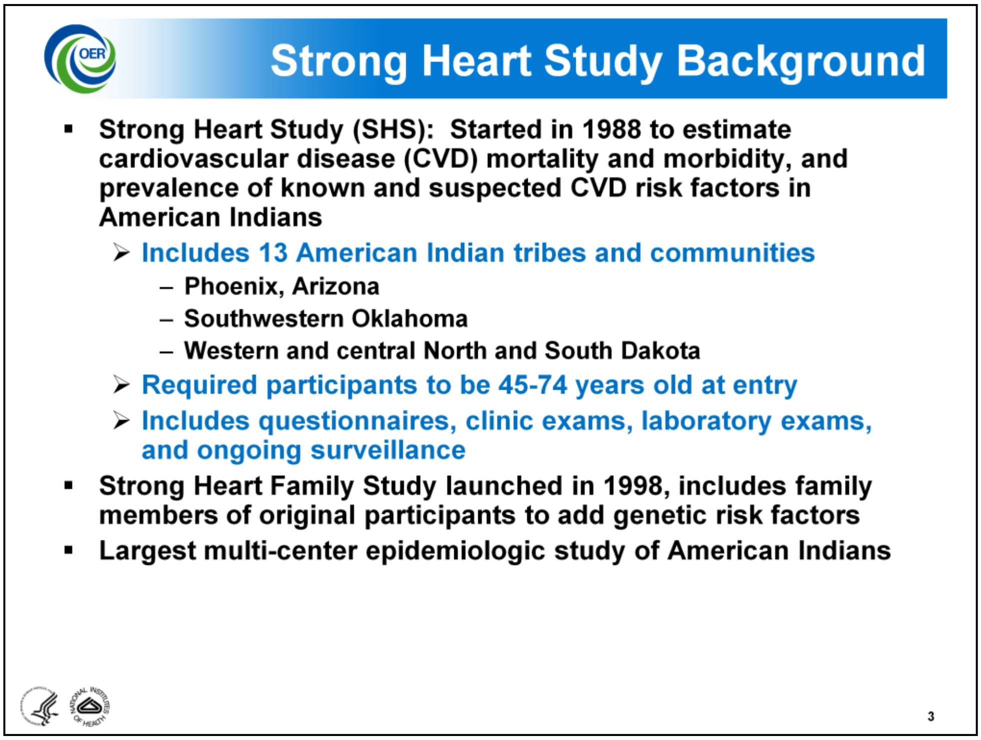
[1]Statistical Analysis and Reporting Branch, Office of Extramural Research,
Office of the Director, National Institutes of Health

[2]Collaborative Health Studies Coordinating Center, University of Washington

[3]Epidemiology Branch, Division of Cardiovascular Sciences,
National Heart, Lung, and Blood Institute, National Institutes of Health

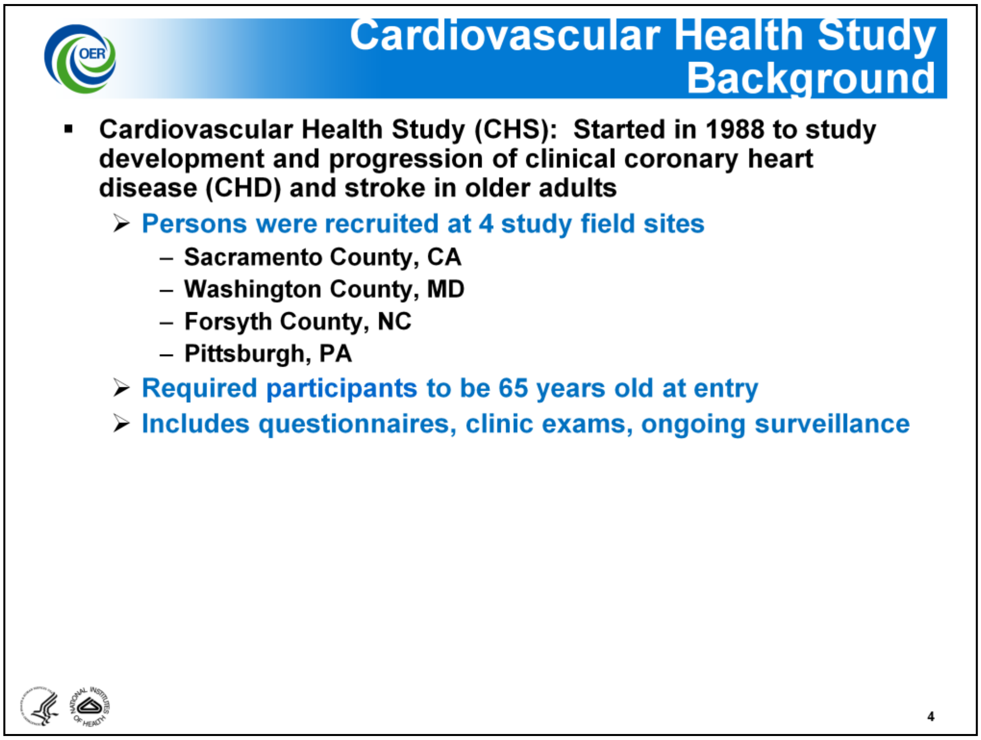
Outline

- Background on NHLBI-funded Cohort Studies
  - ➢ Strong Heart Study
  - ➢ Cardiovascular Health Study
- Research Questions
- Methods
- Descriptive Statistics
- Social Network Analysis
- Conclusions
- Limitations



Strong Heart Study Background

- Strong Heart Study (SHS): Started in 1988 to estimate cardiovascular disease (CVD) mortality and morbidity, and prevalence of known and suspected CVD risk factors in American Indians
  - Includes 13 American Indian tribes and communities
    - Phoenix, Arizona
    - Southwestern Oklahoma
    - Western and central North and South Dakota
  - Required participants to be 45-74 years old at entry
  - Includes questionnaires, clinic exams, laboratory exams, and ongoing surveillance
- Strong Heart Family Study launched in 1998, includes family members of original participants to add genetic risk factors
- Largest multi-center epidemiologic study of American Indians

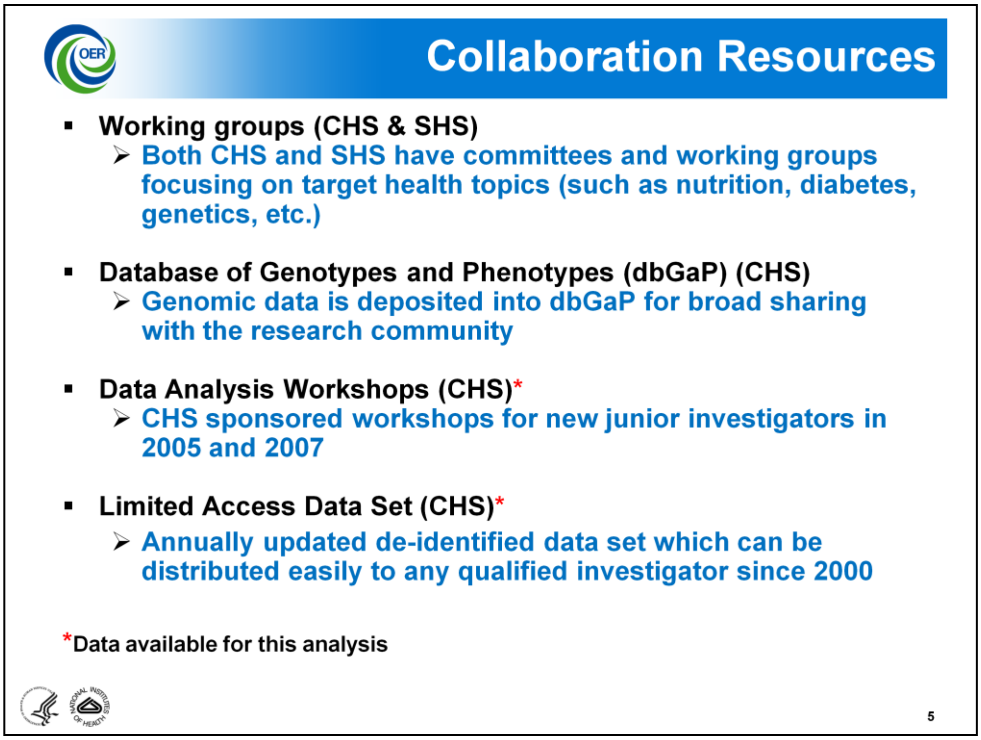


Cardiovascular Health Study Background

- Cardiovascular Health Study (CHS): Started in 1988 to study development and progression of clinical coronary heart disease (CHD) and stroke in older adults
  - ➢ Persons were recruited at 4 study field sites
    - – Sacramento County, CA
    - – Washington County, MD
    - – Forsyth County, NC
    - – Pittsburgh, PA
  - ➢ Required participants to be 65 years old at entry
  - ➢ Includes questionnaires, clinic exams, ongoing surveillance

4

Defining dbGaP



Collaboration Resources

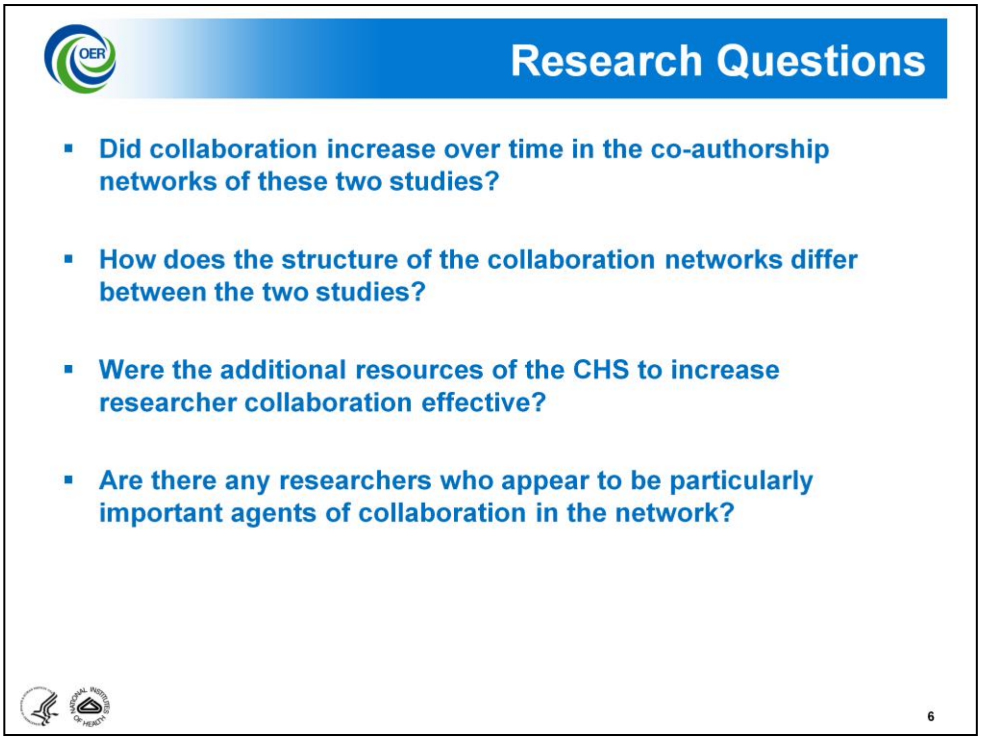
- Working groups (CHS & SHS)
  - Both CHS and SHS have committees and working groups focusing on target health topics (such as nutrition, diabetes, genetics, etc.)
- Database of Genotypes and Phenotypes (dbGaP) (CHS)
  - Genomic data is deposited into dbGaP for broad sharing with the research community
- Data Analysis Workshops (CHS)*
  - CHS sponsored workshops for new junior investigators in 2005 and 2007
- Limited Access Data Set (CHS)*
  - Annually updated de-identified data set which can be distributed easily to any qualified investigator since 2000

*Data available for this analysis

SHS applied for workshop funding, but did not receive.

SHS does not participate in LADS or dbGaP because study participants want more direct control over how their personal data is used.

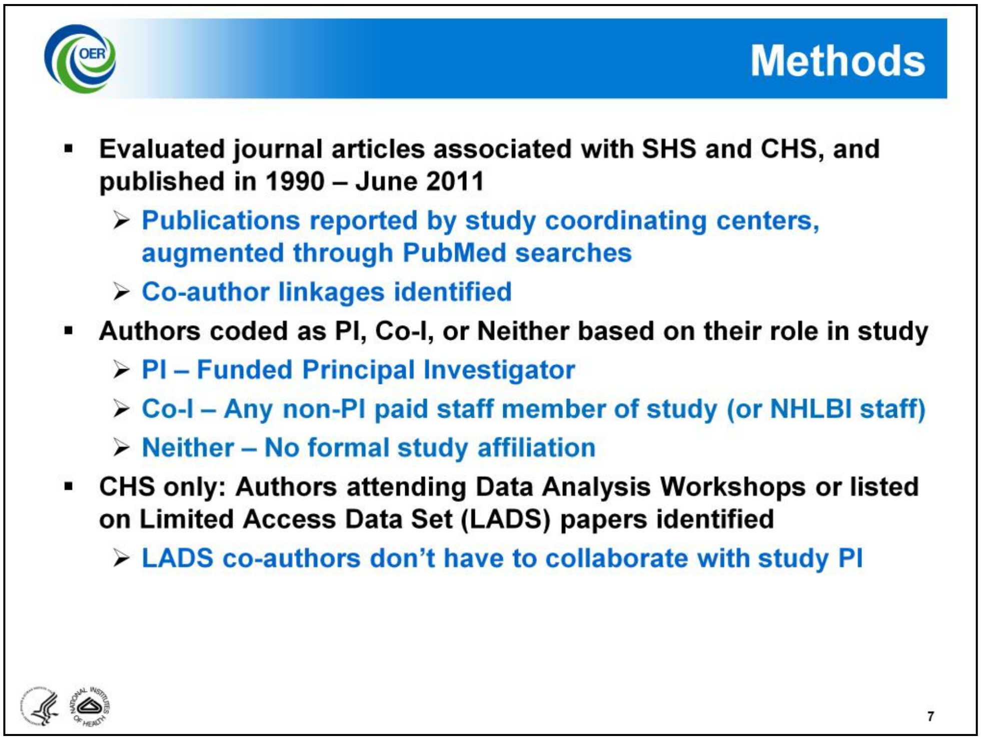
Currently, we only have data to analyze workshop attendance and LADS.



Research Questions

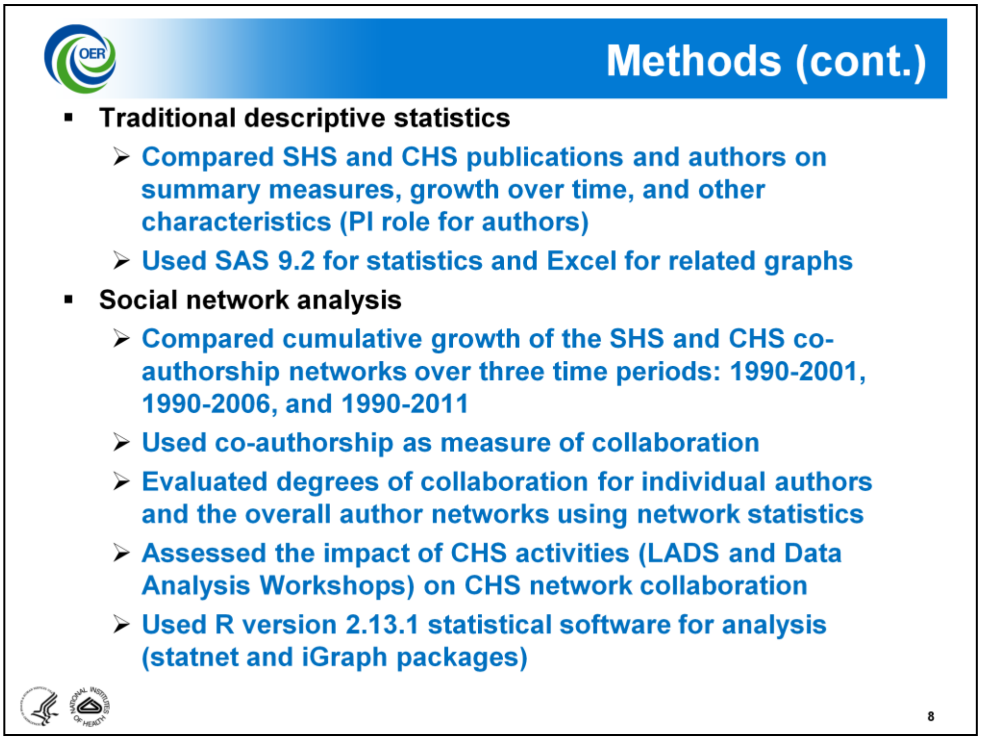
- Did collaboration increase over time in the co-authorship networks of these two studies?
- How does the structure of the collaboration networks differ between the two studies?
- Were the additional resources of the CHS to increase researcher collaboration effective?
- Are there any researchers who appear to be particularly important agents of collaboration in the network?

- Were the additional resources of the CHS to increase researcher collaboration effective?
  - Answered by evaluating Data Analysis Workshops and Limited Access Data Sets



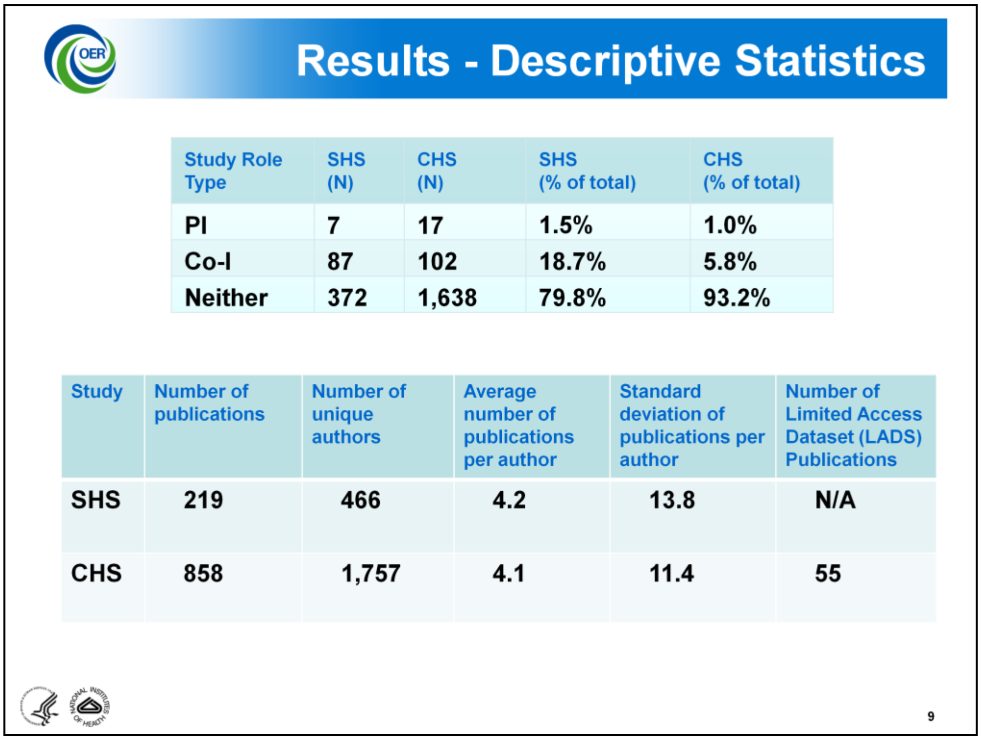
Methods

- Evaluated journal articles associated with SHS and CHS, and published in 1990 – June 2011
  - ➢ Publications reported by study coordinating centers, augmented through PubMed searches
  - ➢ Co-author linkages identified
- Authors coded as PI, Co-I, or Neither based on their role in study
  - ➢ PI – Funded Principal Investigator
  - ➢ Co-I – Any non-PI paid staff member of study (or NHLBI staff)
  - ➢ Neither – No formal study affiliation
- CHS only: Authors attending Data Analysis Workshops or listed on Limited Access Data Set (LADS) papers identified
  - ➢ LADS co-authors don't have to collaborate with study PI



Methods (cont.)

- Traditional descriptive statistics
  - Compared SHS and CHS publications and authors on summary measures, growth over time, and other characteristics (PI role for authors)
  - Used SAS 9.2 for statistics and Excel for related graphs
- Social network analysis
  - Compared cumulative growth of the SHS and CHS co-authorship networks over three time periods: 1990-2001, 1990-2006, and 1990-2011
  - Used co-authorship as measure of collaboration
  - Evaluated degrees of collaboration for individual authors and the overall author networks using network statistics
  - Assessed the impact of CHS activities (LADS and Data Analysis Workshops) on CHS network collaboration
  - Used R version 2.13.1 statistical software for analysis (statnet and iGraph packages)



Results - Descriptive Statistics

Study Role Type	SHS (N)	CHS (N)	SHS (% of total)	CHS (% of total)
PI	7	17	1.5%	1.0%
Co-I	87	102	18.7%	5.8%
Neither	372	1,638	79.8%	93.2%

Study	Number of publications	Number of unique authors	Average number of publications per author	Standard deviation of publications per author	Number of Limited Access Dataset (LADS) Publications
SHS	219	466	4.2	13.8	N/A
CHS	858	1,757	4.1	11.4	55

Here we see that the SHS had fewer PIs and Co-Is than the CHS. However, PIs and Co-Is represented a higher proportion of authors in the SHS compared to the CHS, which had proportionately more unaffiliated authors.

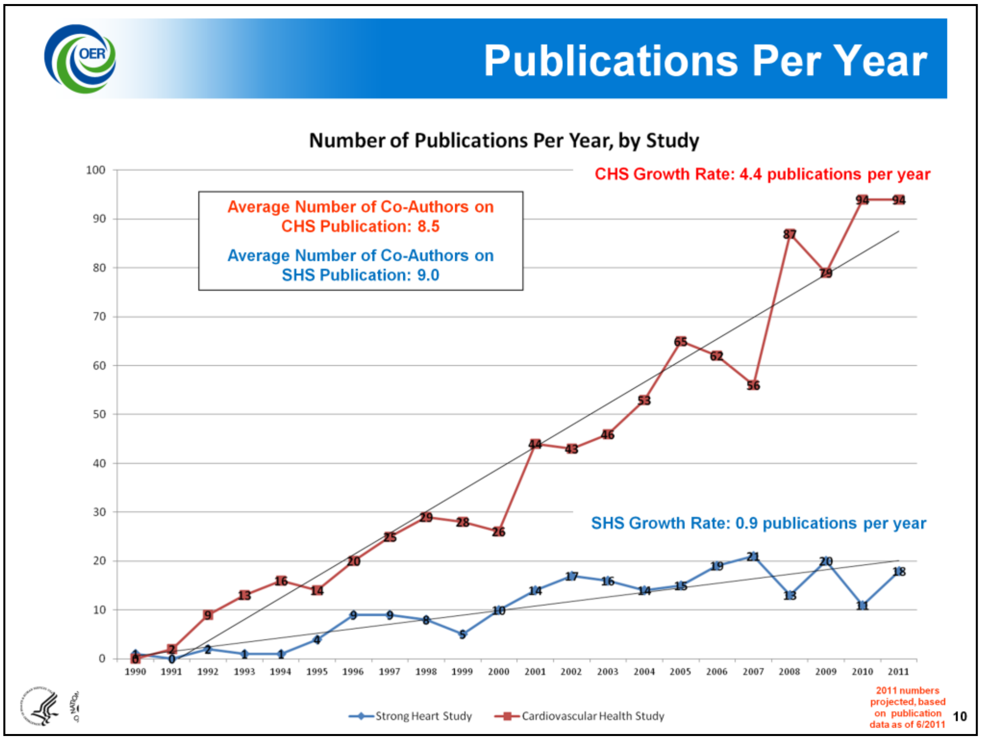
SHS had fewer total unique authors and thus, publications, than CHS [DESCRIBE]. Despite this, the average number of publications per author was nearly the same for both studies – around 4 per author. The CHS had 55 LADS publications.

There were 219 SHS publications, published by a total of 466 unique co-authors from 1990-2011

- Average number of publications per author was 4.2 (standard deviation - 13.8)

There were 858 CHS publications, published by a total of 1,757 unique co-authors from 1990-2011

- Average number of publications per author was 4.1 (standard deviation 11.4)
- 55 publications used the Limited Access Dataset



Publications Per Year

Number of Publications Per Year, by Study

CHS Growth Rate: 4.4 publications per year

Average Number of Co-Authors on CHS Publication: 8.5

Average Number of Co-Authors on SHS Publication: 9.0

SHS Growth Rate: 0.9 publications per year

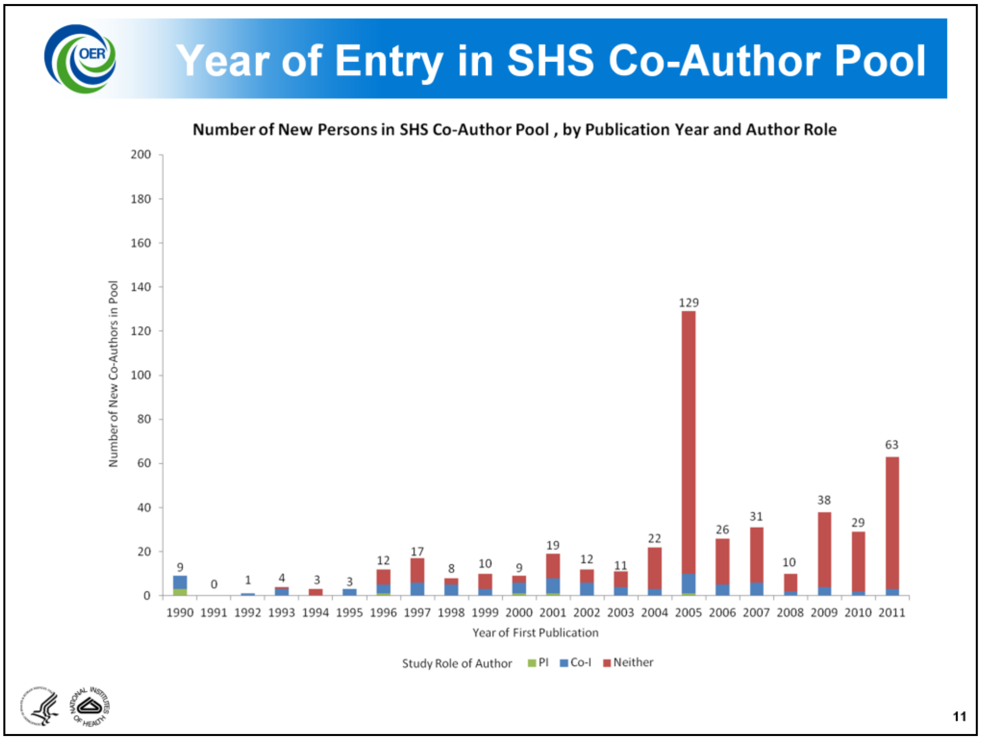
2011 numbers projected, based on publication data as of 6/2011

10

Both studies had similar average numbers of co-authors per publication over the entire evaluation period (1990-2011)

While both studies experienced growth in the number of publication per year, the CHS grew more rapidly than did SHS (4.4 v. 0.9 publications per year, respectively).

If JAMA paper removed from SHS, then average number of SHS authors goes down to 8.5

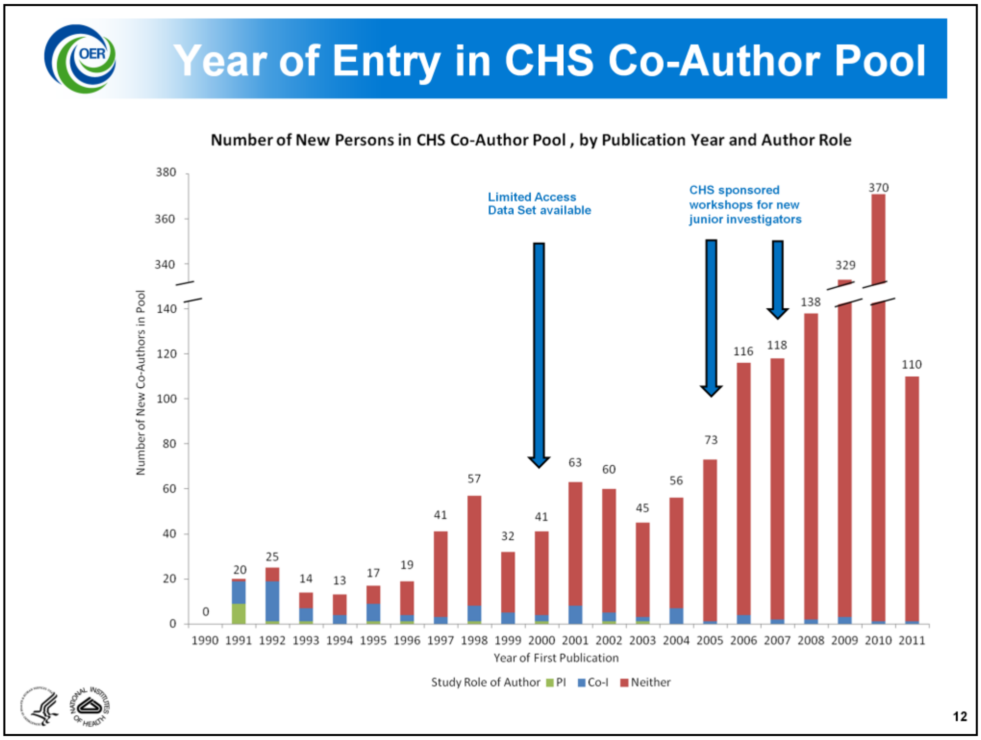


Year of Entry in SHS Co-Author Pool

Here we see the year of entry of co-authors into the network for the SHS study. Each author is represented only once on the graph.

In conjunction with the growth in publications, the number of new authors in each year grew as well. In the early years, PIs and Co-Is dominated (green and blue). In the later years (2000 and beyond) much of this growth was from co-authors who were NOT PIs or Co-Is (shown in red).

One study with large number of co-authors accounted for the large spike of new authors entering the network seen in 2005.

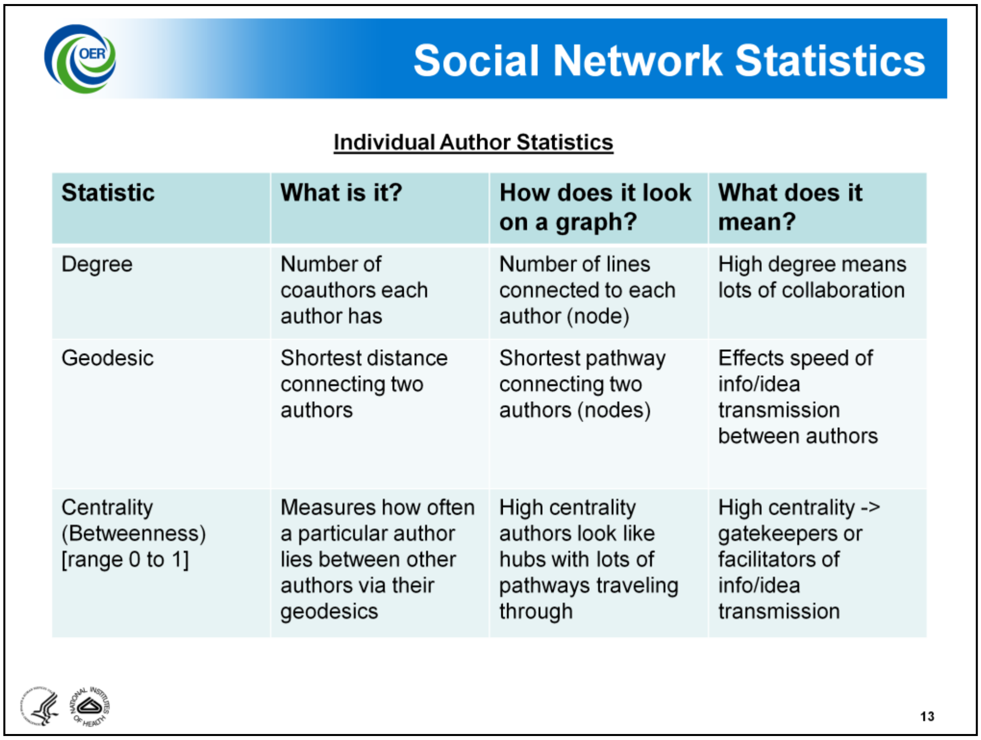


Year of Entry in CHS Co-Author Pool

Here we see the year of entry of co-authors for the CHS study.

Again, PIs and Co-Is dominated in the early years. In conjunction with the sharp growth in the study publications starting around 2000, number of new co-authors grew very rapidly, with non-study-affiliated authors increasing the most.

The number of new authors entering the network each year increased after the Limited Access Data Sets became available in 2000 and the Data Analysis Workshops were held in 2005 and 2007.

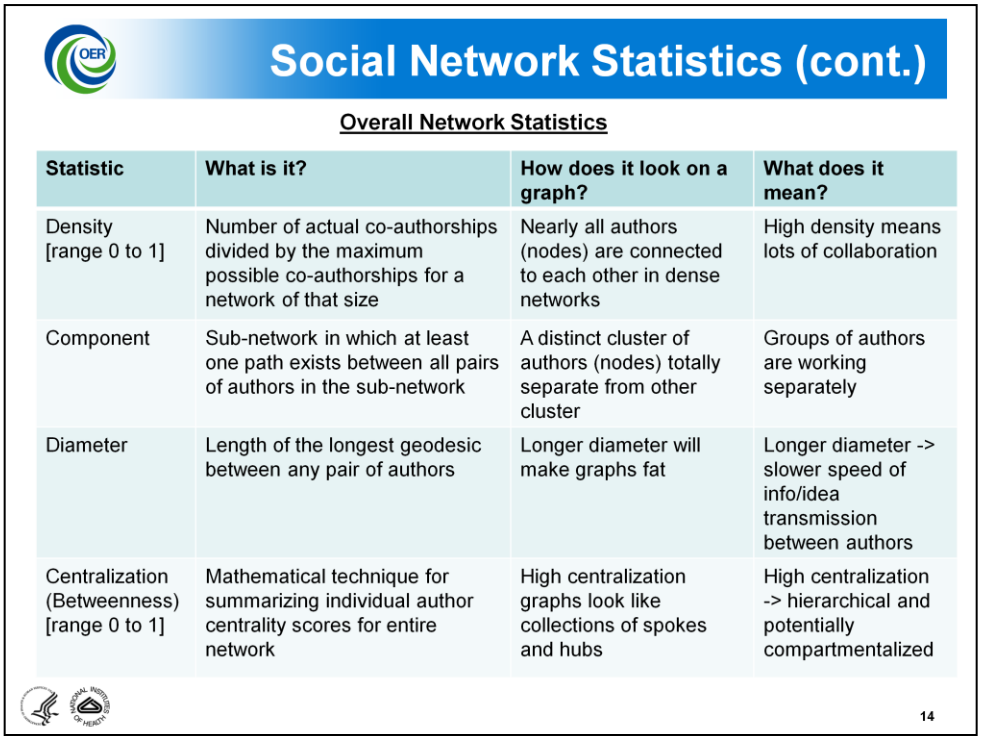


Social Network Statistics

Individual Author Statistics

Statistic	What is it?	How does it look on a graph?	What does it mean?
Degree	Number of coauthors each author has	Number of lines connected to each author (node)	High degree means lots of collaboration
Geodesic	Shortest distance connecting two authors	Shortest pathway connecting two authors (nodes)	Effects speed of info/idea transmission between authors
Centrality (Betweenness) [range 0 to 1]	Measures how often a particular author lies between other authors via their geodesics	High centrality authors look like hubs with lots of pathways traveling through	High centrality -> gatekeepers or facilitators of info/idea transmission

These definitions will be followed by specific examples

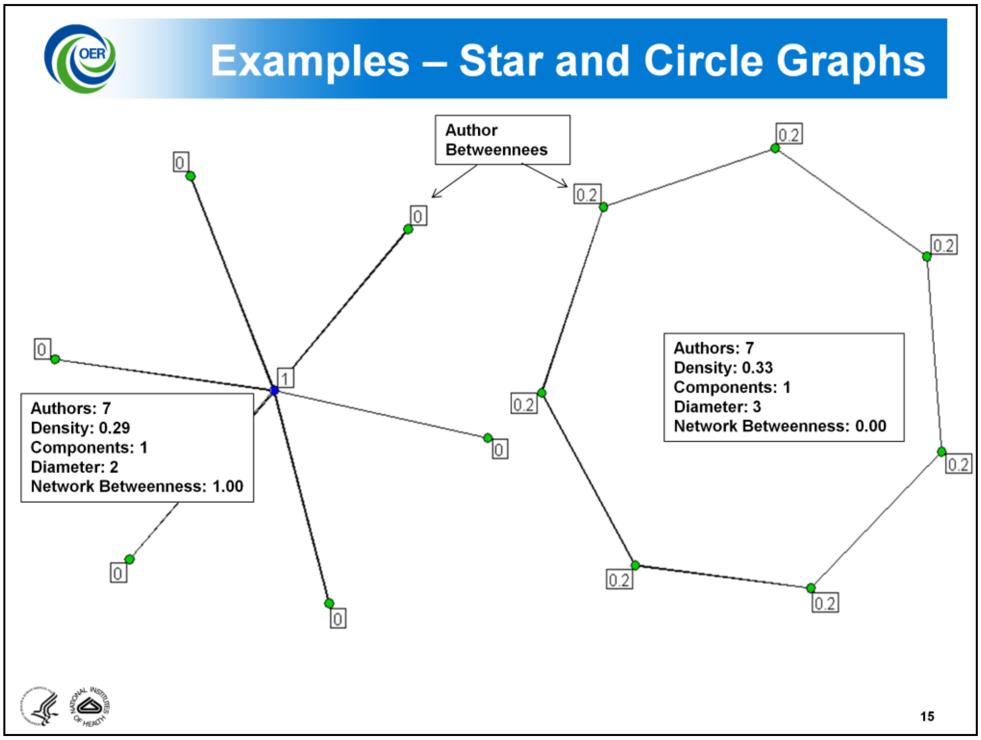


Social Network Statistics (cont.)

Overall Network Statistics

Statistic	What is it?	How does it look on a graph?	What does it mean?
Density [range 0 to 1]	Number of actual co-authorships divided by the maximum possible co-authorships for a network of that size	Nearly all authors (nodes) are connected to each other in dense networks	High density means lots of collaboration
Component	Sub-network in which at least one path exists between all pairs of authors in the sub-network	A distinct cluster of authors (nodes) totally separate from other cluster	Groups of authors are working separately
Diameter	Length of the longest geodesic between any pair of authors	Longer diameter will make graphs fat	Longer diameter -> slower speed of info/idea transmission between authors
Centralization (Betweenness) [range 0 to 1]	Mathematical technique for summarizing individual author centrality scores for entire network	High centralization graphs look like collections of spokes and hubs	High centralization -> hierarchical and potentially compartmentalized

These definitions will be followed by specific examples



Examples – Star and Circle Graphs

Author Betweennees

Authors: 7
Density: 0.29
Components: 1
Diameter: 2
Network Betweenness: 1.00

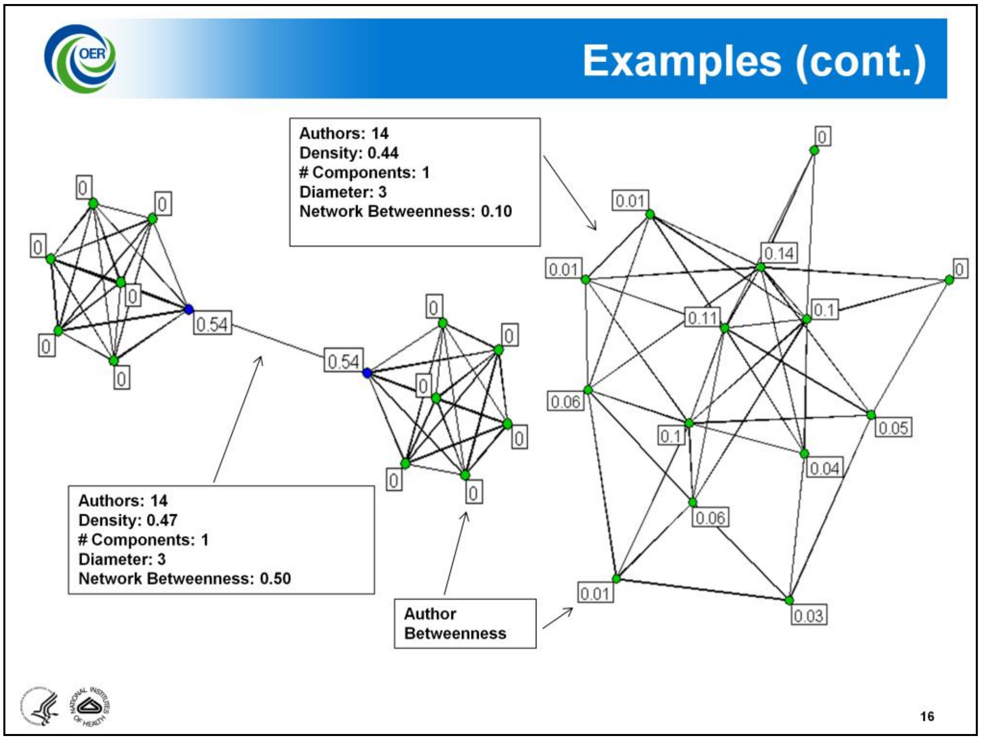
Authors: 7
Density: 0.33
Components: 1
Diameter: 3
Network Betweenness: 0.00

Both graphs have the same number of authors, 7. The numbers next to each author (or node) represent that person's betweeness score.

First, star graph. The central author in blue has a betweenness score of 1, because he has published six papers, one with every other author and none of the other authors have published with each other.

For the network as a whole, the density is 0.29 (about 30% of the possible connections have been made), 1 component, and the diameter is 2 (it takes two steps to go from one end of the network to the other). The overall graph also is maximally centralized, with a network betweenness score of 1. You can't get any more centralized than this graph. It is very hierarchical structure.

Circle graph is the exact opposite. There is no centralized author – each author has the exact same betweenness score and thus the network betweenness is 0. Each author has written two papers, with one co-author on each. This is a very diffuse structure, and it has no hierarchy.



Examples (cont.)

Authors: 14
Density: 0.44
Components: 1
Diameter: 3
Network Betweenness: 0.10

Authors: 14
Density: 0.47
Components: 1
Diameter: 3
Network Betweenness: 0.50

Author Betweenness

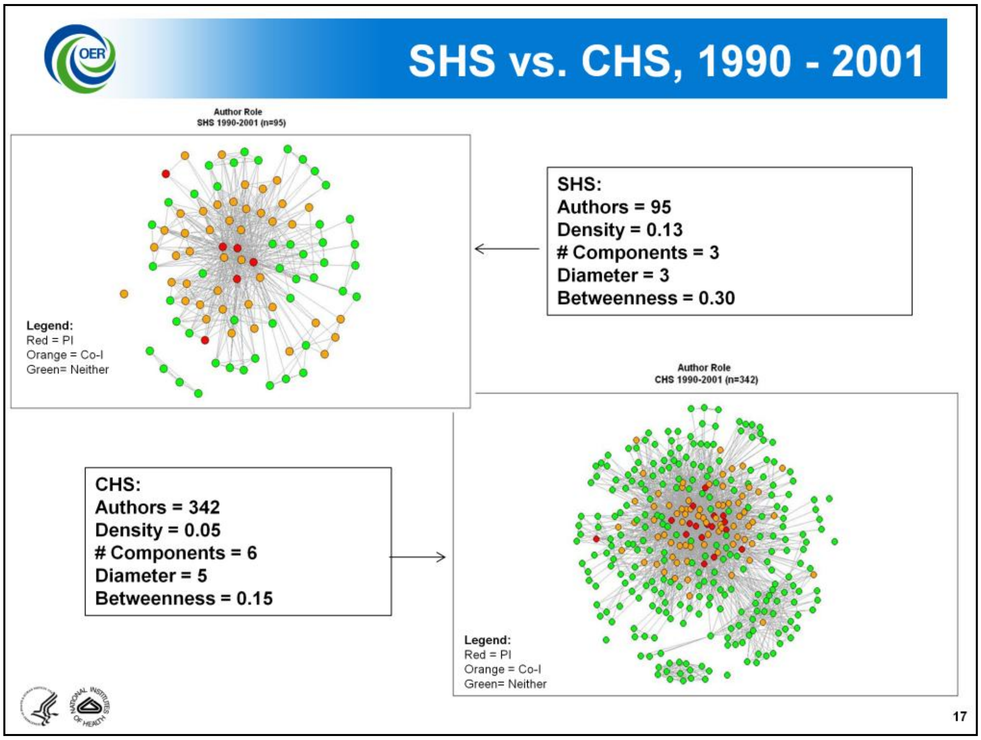
Again, networks of the same size (14 authors), with about the same density, # components, and diameter.

Start with the dumbbell graph on the left. The betweenness scores of the authors on each end of the dumbbell are zero, so they don't lie between any two authors in the network. The blue authors on the other hand have very high betweenness scores, because they form the only pathway from one end of the dumbbell to the other. This causes the overall network betweenness to be quite high, 0.5. The dumbbell graph could have arisen by the authors on the left dumbbell all collaborating on a single publication together, the authors on the right dumbbell collaborating on a single publication together, and the two blue authors collaborating on a publication together. Again, this is clearly a hierarchical structure.

The graph on the right is much more loosely organized. There is very little betweenneess power for any single author, so the individual author betweenness scores are low as well as the overall network betweenness.

It is not entirely clear which network is a better model for collaboration. The dumbbell graph could be very efficient for disseminating info/ideas or it could be virtually segregated into two different networks – it really depends on the behavior of the blue authors.

Now we turn to our actual analyses of the CHS and SHS.



SHS vs. CHS, 1990 - 2001

Author Role
SHS 1990-2001 (n=95)

Legend:
Red = PI
Orange = Co-I
Green= Neither

SHS:
Authors = 95
Density = 0.13
# Components = 3
Diameter = 3
Betweenness = 0.30

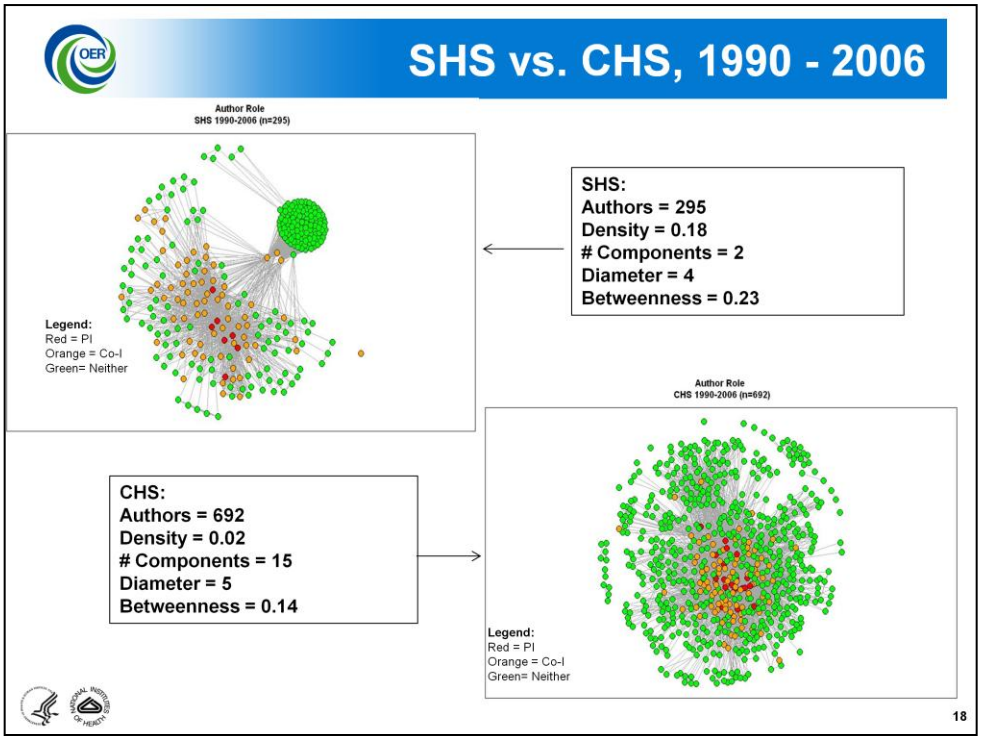
CHS:
Authors = 342
Density = 0.05
# Components = 6
Diameter = 5
Betweenness = 0.15

Author Role
CHS 1990-2001 (n=342)

Legend:
Red = PI
Orange = Co-I
Green= Neither

Here we compare the SHS and CHS in the earliest time period, 1990-2001. In these early years, especially for the SHS, we see a relatively large preponderance of PIs and Co-Is (shown in red and orange) who were key to launching and managing the studies, compared to other authors (shown in green).

CHS is larger (or has more authors), is less dense (though this is a partially a function of largeness – it is inherently difficult for all authors in large networks to co-author with one another), and has a larger diameter and more components (or more groups of authors publishing separately from the others). SHS is characterized by greater centralization (betweenneess is twice that of CHS). You can see it has more of a hub and spoke look.



SHS vs. CHS, 1990 - 2006

Author Role
SHS 1990-2006 (n=295)

Legend:
Red = PI
Orange = Co-I
Green= Neither

SHS:
Authors = 295
Density = 0.18
# Components = 2
Diameter = 4
Betweenness = 0.23

CHS:
Authors = 692
Density = 0.02
# Components = 15
Diameter = 5
Betweenness = 0.14

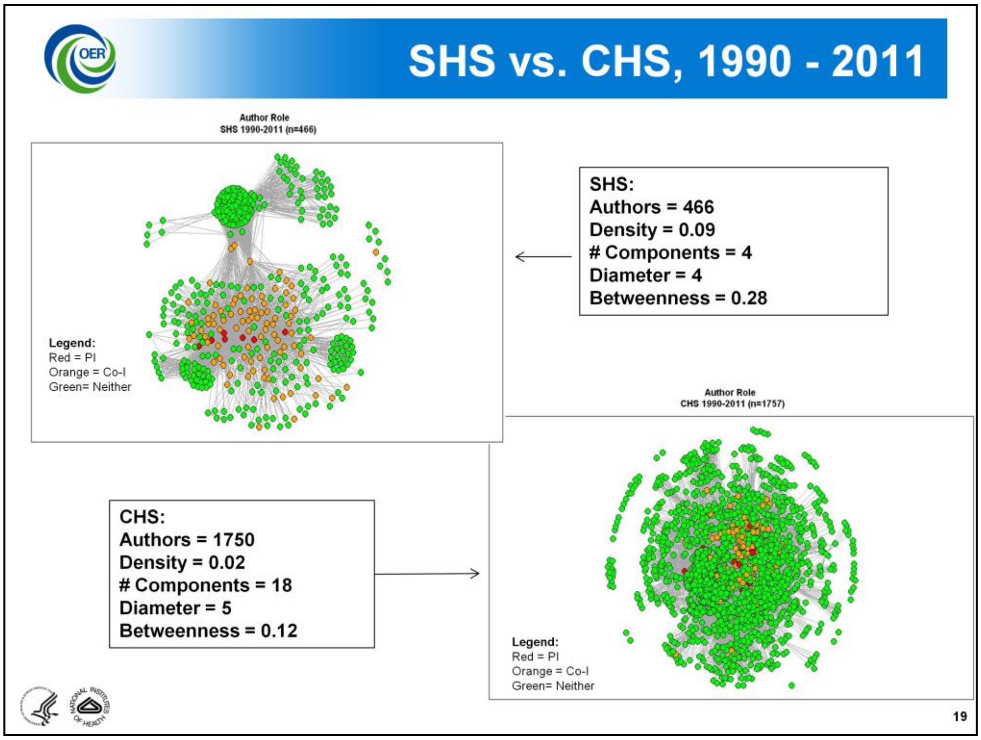
Author Role
CHS 1990-2006 (n=692)

Legend:
Red = PI
Orange = Co-I
Green= Neither

Here we compare the networks from 1990 through 2006. The statistical trends from the earlier period continue.

SHS: JAMA paper PMID 16219884 in 2005 with 115 unique authors, creating a large “tumor” in the graph that increases the network density.

In CHS, the PIs and co-Is are increasingly surrounded by other authors, those with no formal affiliation to the study, shown in green.



SHS vs. CHS, 1990 - 2011

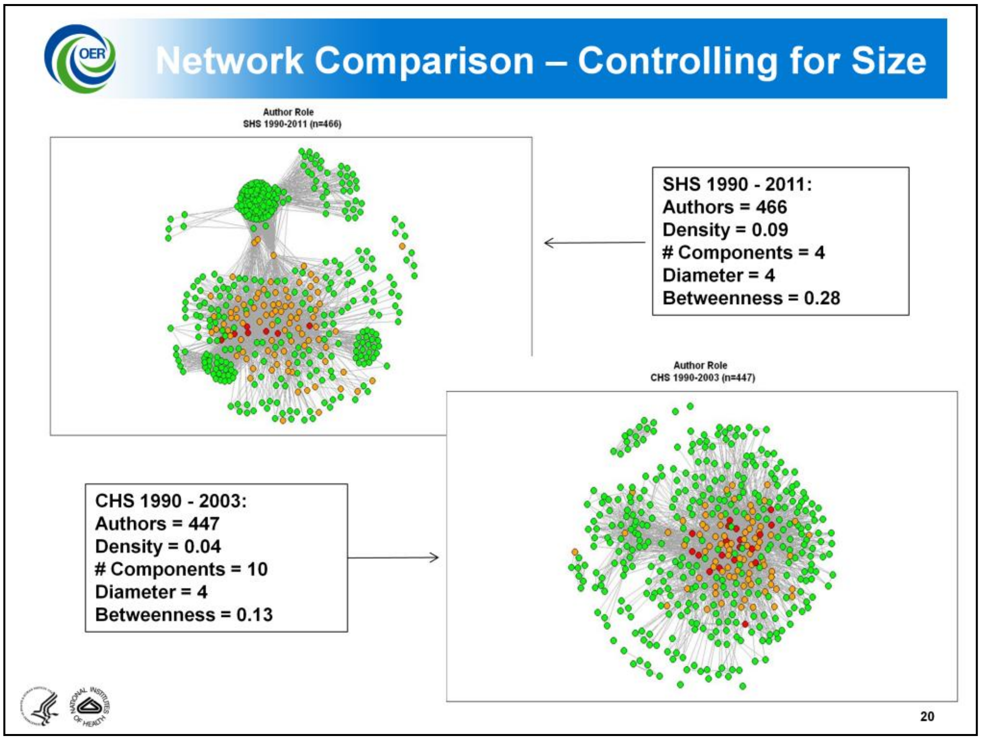
SHS:
Authors = 466
Density = 0.09
# Components = 4
Diameter = 4
Betweenness = 0.28

CHS:
Authors = 1750
Density = 0.02
# Components = 18
Diameter = 5
Betweenness = 0.12

Here we compare both networks for the entire evaluation period, 1990-2011.

Compared to the SHS, the CHS has many more authors, who co-author with a smaller fraction of all authors in the network (measured by a lower density). CHS authors are more distant from each other (as indicated by a longer diameter) and have more groups publishing separately from the main network (in other words, have more components).

SHS has a higher betweenness, so it is more centralized with certain co-authors serving as key facilitators of collaboration.



Network Comparison – Controlling for Size

Author Role
SHS 1990-2011 (n=466)

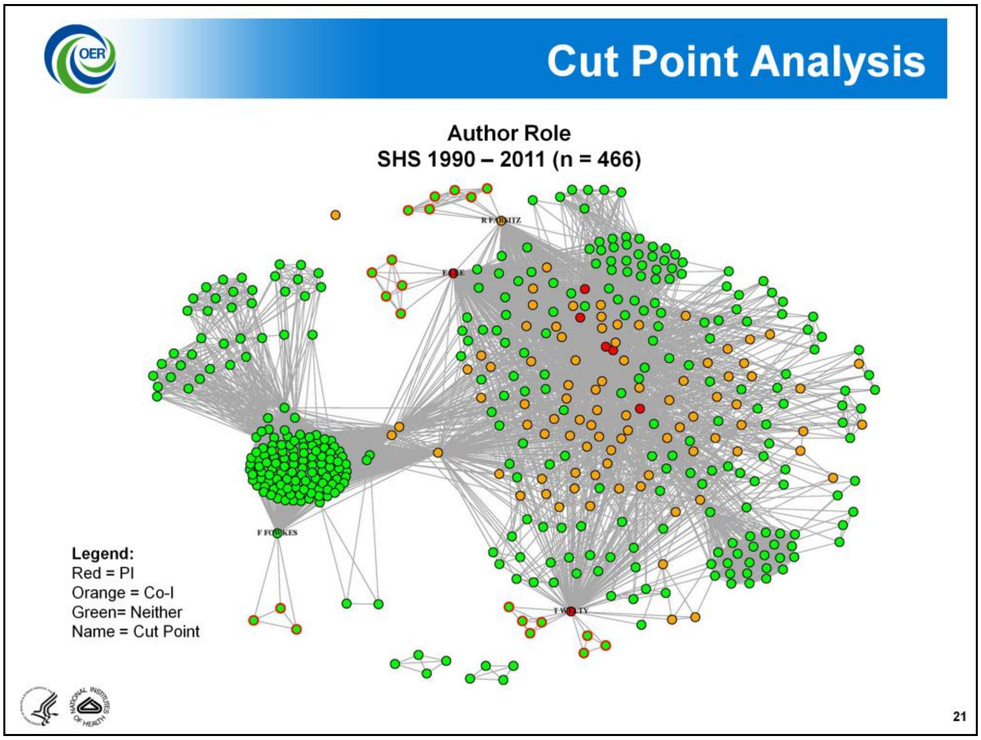
SHS 1990 - 2011:
Authors = 466
Density = 0.09
# Components = 4
Diameter = 4
Betweenness = 0.28

Author Role
CHS 1990-2003 (n=447)

CHS 1990 - 2003:
Authors = 447
Density = 0.04
# Components = 10
Diameter = 4
Betweenness = 0.13

20

This slide shows the two networks at different points in time, but where the size of the networks – or number of authors -- are nearly the same. Now, the size-dependent statistics, like density, are more easily comparable. The same story line holds up. CHS is less dense, meaning there are fewer authors publishing with each other, with more components, i.e., more authors publishing separately from the others. SHS is more centralized and hierarchical, with key facilitators playing an influential role.



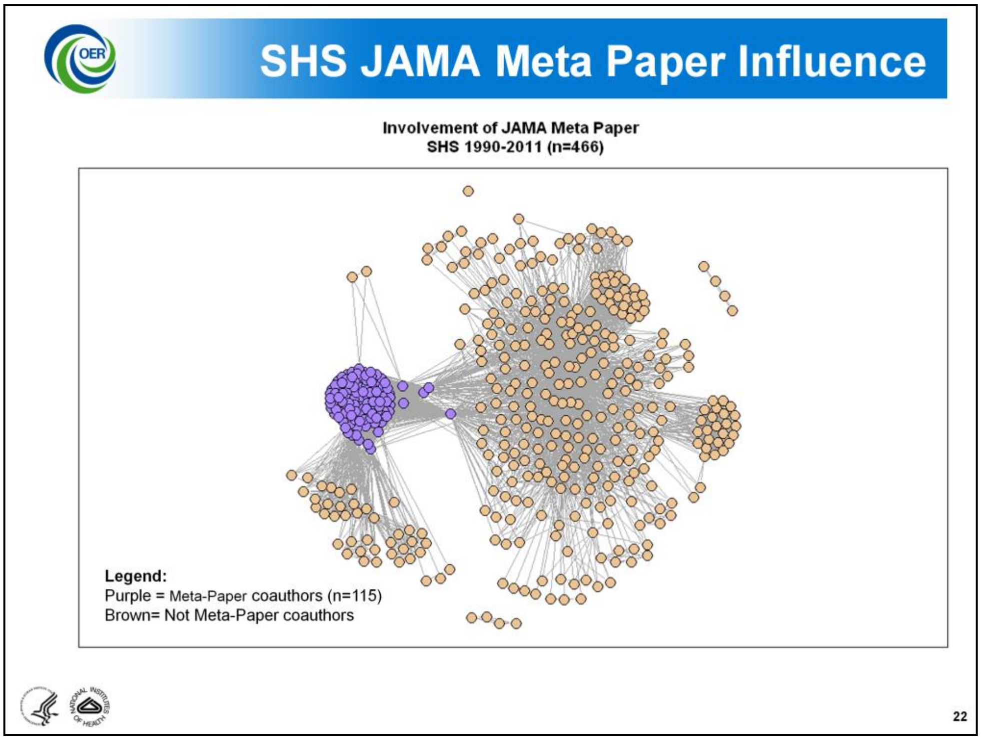
Cut Point Analysis

Author Role
SHS 1990 – 2011 (n = 466)

Legend:
Red = PI
Orange = Co-I
Green= Neither
Name = Cut Point

21

A closer look at the SHS network. Cut points are authors who, if removed, would cause the network to split up into more components – so they are clearly important authors in the network. However, many of the most valuable authors, in terms of betweenness, are not cut points. We took the standardized individual betweenness scores from the SHS and CHS networks combined and identified the top 10 facilitators of co-authorship. Only 2 of these top 10 were cut point authors (ranked 4 and 9). Seven of the top ten belong to SHS.



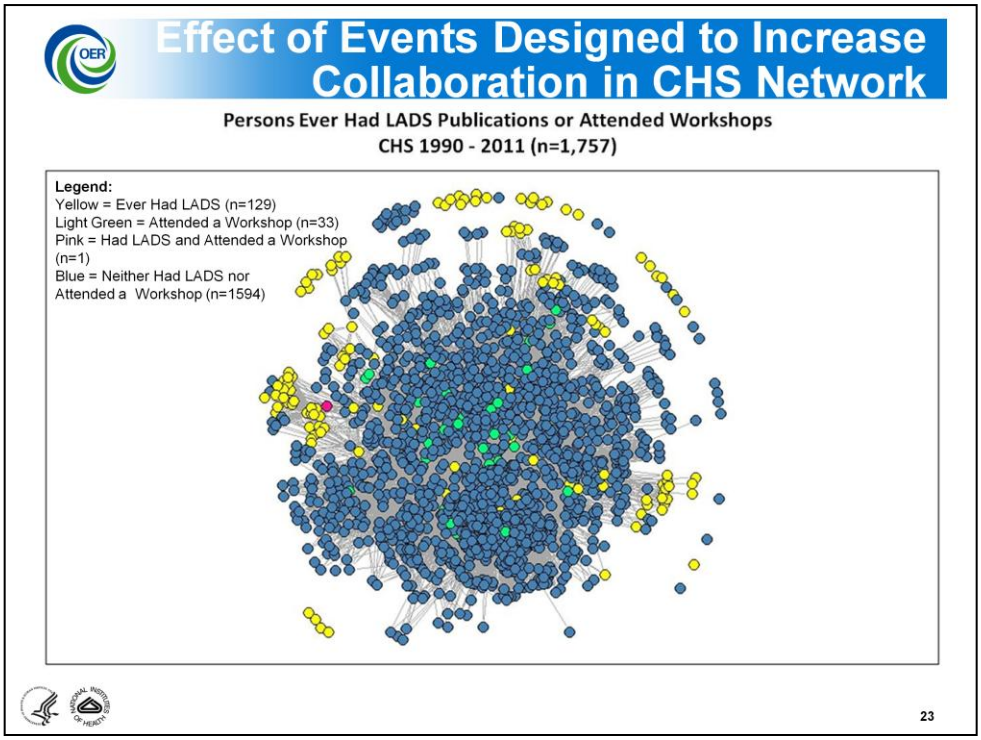
SHS JAMA Meta Paper Influence

Involvement of JAMA Meta Paper
SHS 1990-2011 (n=466)

Legend:
Purple = Meta-Paper coauthors (n=115)
Brown= Not Meta-Paper coauthors

22

JAMA paper PMID 16219884 in 2005 with 115 unique authors, increases density of overall network (somewhat artificially).



Effect of Events Designed to Increase Collaboration in CHS Network

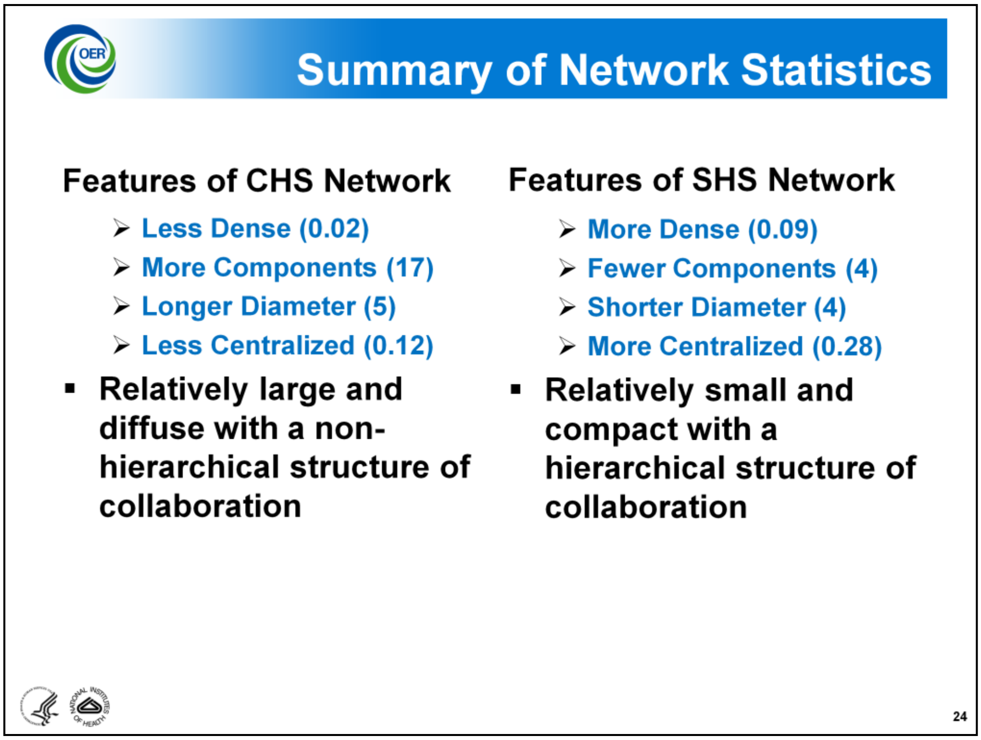
Persons Ever Had LADS Publications or Attended Workshops
CHS 1990 - 2011 (n=1,757)

Legend:
Yellow = Ever Had LADS (n=129)
Light Green = Attended a Workshop (n=33)
Pink = Had LADS and Attended a Workshop (n=1)
Blue = Neither Had LADS nor Attended a Workshop (n=1594)

23

163 of the 1,757 co-authors – or 9% of the total – who entered the CHS network participated in a Limited Access Data Set (LADS) publication or attended a Data Analysis Workshop.

Recall that authors using LADS data are not required to collaborate with CHS PIs. We see that many of the clusters of co-authors completely separated from the main network are indeed associated with LADS papers (shown in yellow). Data Analysis Workshop attendees (shown in light green) have managed to become relatively centralized, or important facilitators of collaboration, in a short period of time (since 2005).



Summary of Network Statistics

Features of CHS Network

- Less Dense (0.02)
- More Components (17)
- Longer Diameter (5)
- Less Centralized (0.12)

- Relatively large and diffuse with a non-hierarchical structure of collaboration

Features of SHS Network

- More Dense (0.09)
- Fewer Components (4)
- Shorter Diameter (4)
- More Centralized (0.28)

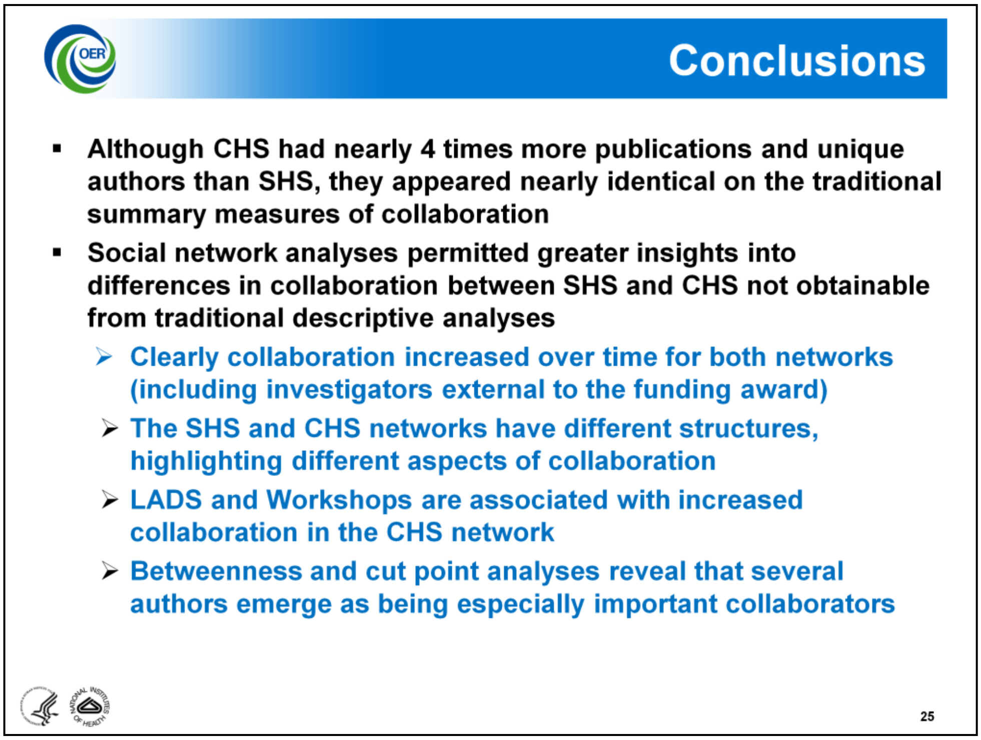
- Relatively small and compact with a hierarchical structure of collaboration

Density, our purest measure of collaboration, is higher for SHS. Density is indirectly a function of size, but even comparing similarly sized graphs, SHS is more dense.

CHS has more components, mainly a result of the availability of LADS, which allow independent investigators to work outside the network.

CHS has a longer diameter (also an indirect function of size), meaning that more steps have to be traversed for info/ideas to cross from one extreme of the network to the other.

SHS is more centralized (in terms of betweenness), meaning that certain authors are especially valuable to the network in terms of serving as points of connection between different sections of the network (i.e., gatekeepers of facilitators). This could make the network more efficient or compartmentalized, depending on the behavior of the gatekeepers.

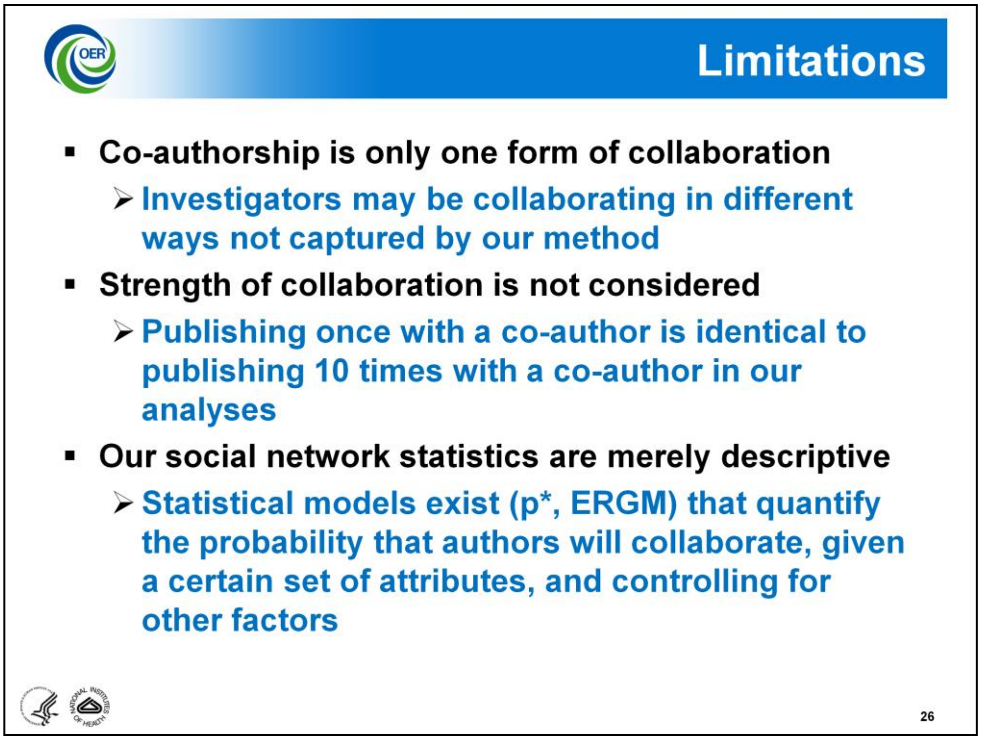


Conclusions

- Although CHS had nearly 4 times more publications and unique authors than SHS, they appeared nearly identical on the traditional summary measures of collaboration
- Social network analyses permitted greater insights into differences in collaboration between SHS and CHS not obtainable from traditional descriptive analyses
  - ➢ Clearly collaboration increased over time for both networks (including investigators external to the funding award)
  - ➢ The SHS and CHS networks have different structures, highlighting different aspects of collaboration
  - ➢ LADS and Workshops are associated with increased collaboration in the CHS network
  - ➢ Betweenness and cut point analyses reveal that several authors emerge as being especially important collaborators

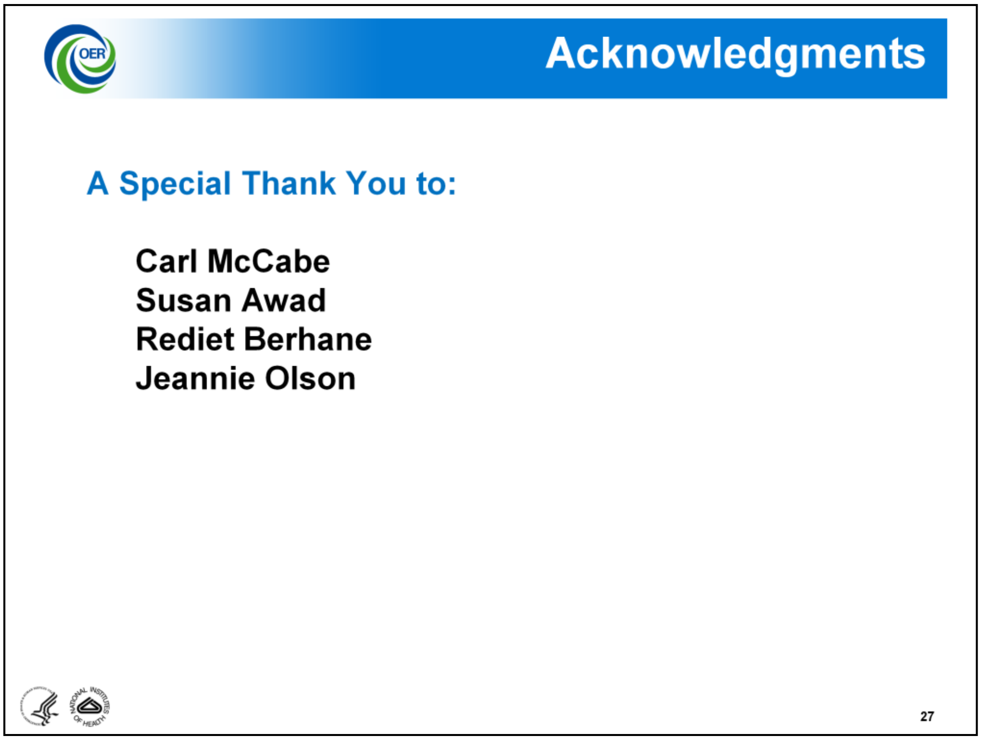
25

Although CHS had 3.9 times more publications and 3.8 times more unique authors than SHS (854/219 and 1,752/466, respectively), they appeared nearly identical on the summary measures of collaboration: average number of publications per author and average number of co-authors per publication



Limitations

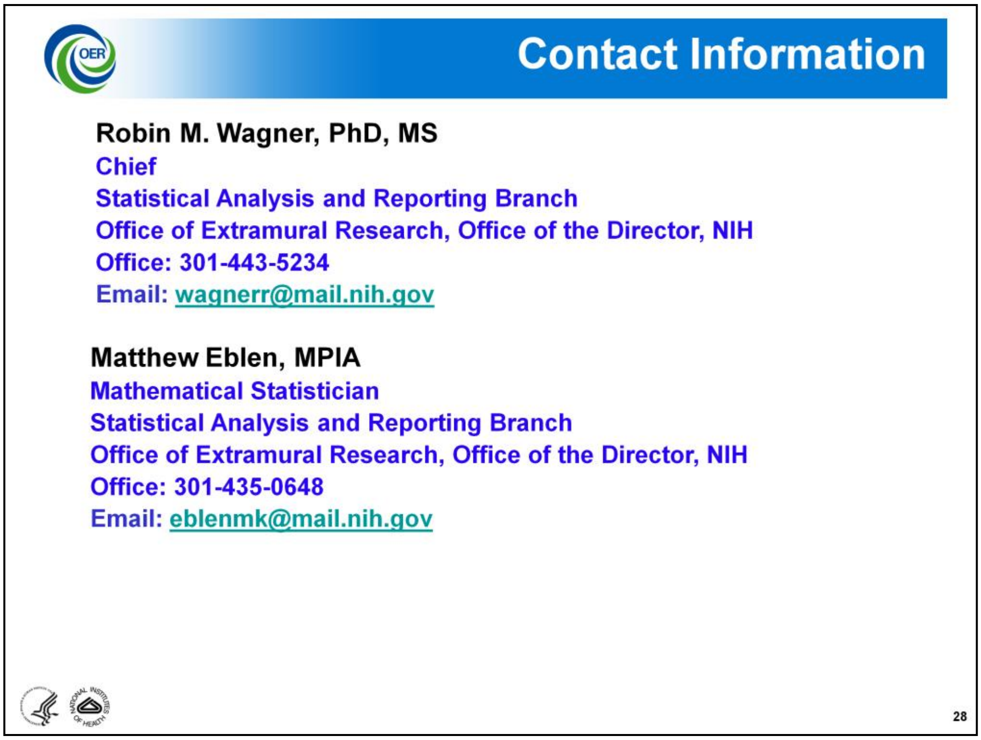
- Co-authorship is only one form of collaboration
  - Investigators may be collaborating in different ways not captured by our method
- Strength of collaboration is not considered
  - Publishing once with a co-author is identical to publishing 10 times with a co-author in our analyses
- Our social network statistics are merely descriptive
  - Statistical models exist (p*, ERGM) that quantify the probability that authors will collaborate, given a certain set of attributes, and controlling for other factors



Acknowledgments

A Special Thank You to:

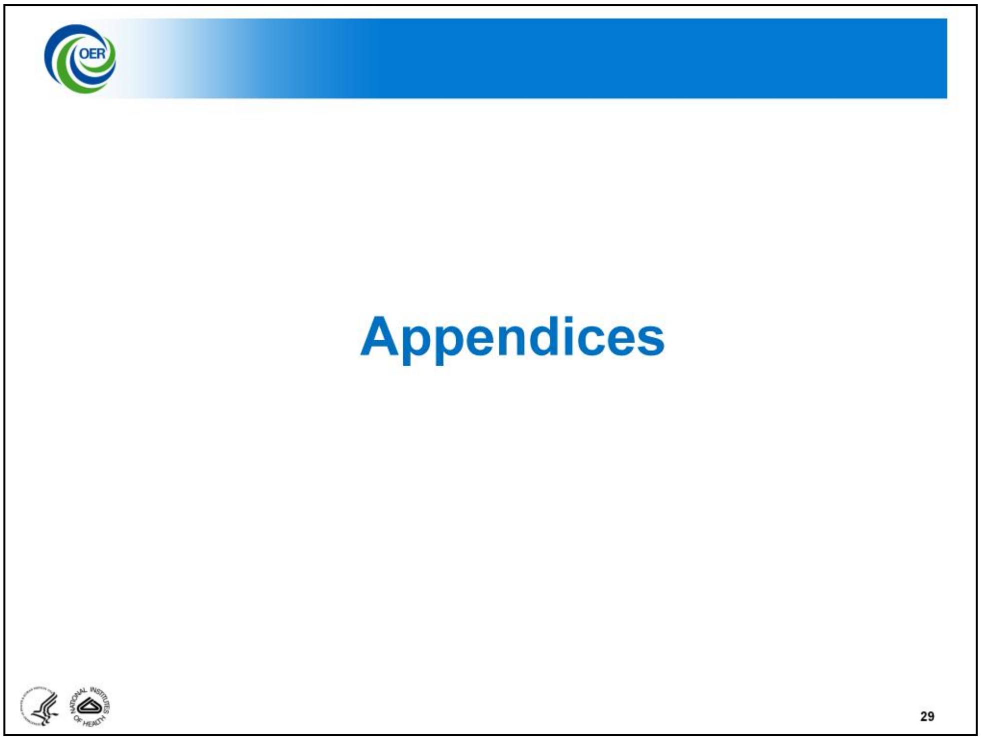
Carl McCabe
Susan Awad
Rediet Berhane
Jeannie Olson



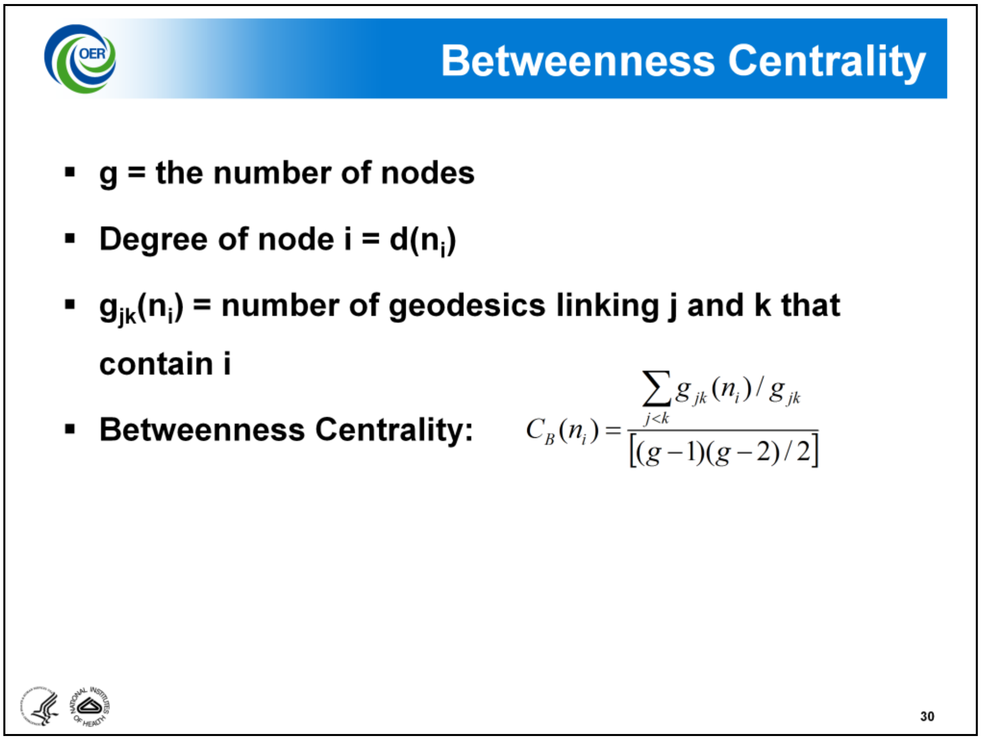
Contact Information

Robin M. Wagner, PhD, MS
Chief
Statistical Analysis and Reporting Branch
Office of Extramural Research, Office of the Director, NIH
Office: 301-443-5234
Email: wagnerr@mail.nih.gov

Matthew Eblen, MPIA
Mathematical Statistician
Statistical Analysis and Reporting Branch
Office of Extramural Research, Office of the Director, NIH
Office: 301-435-0648
Email: eblenmk@mail.nih.gov

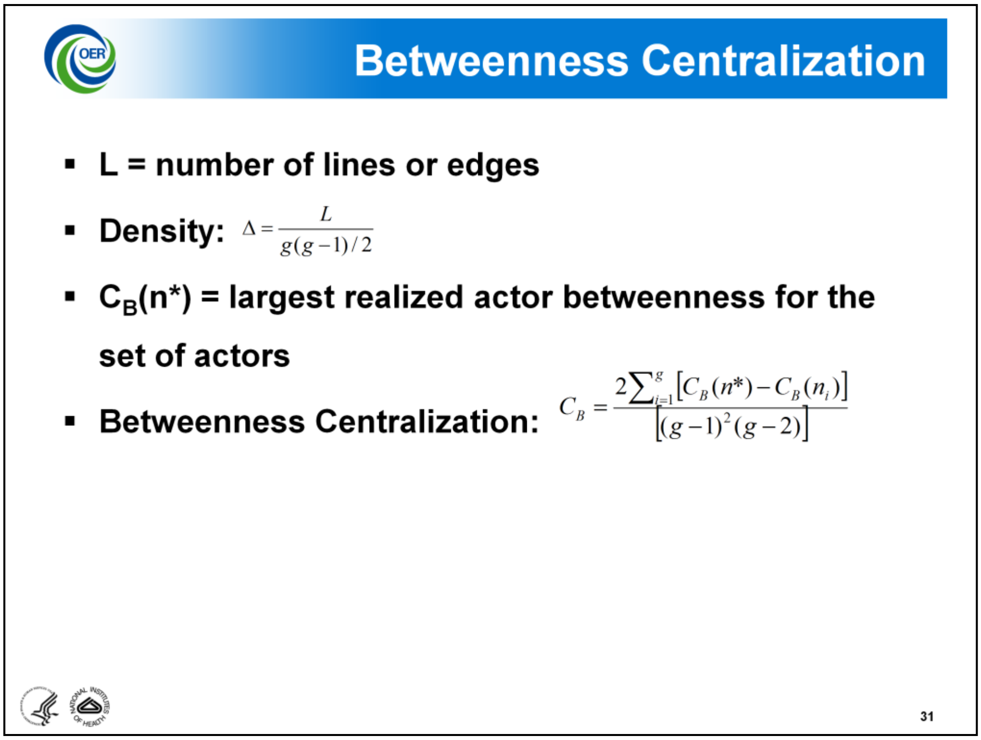


Appendices



Betweenness Centrality

- g = the number of nodes
- Degree of node i = d(n_i)
- $g_{jk}(n_i)$ = number of geodesics linking j and k that contain i
- Betweenness Centrality: $$C_B(n_i) = \frac{\sum_{j<k} g_{jk}(n_i) / g_{jk}}{[(g-1)(g-2)/2]}$$

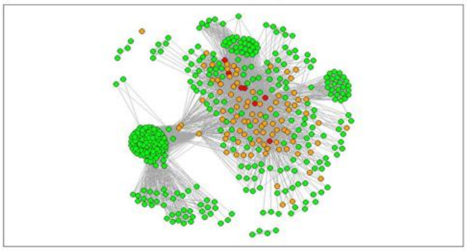


Betweenness Centralization

- L = number of lines or edges
- Density: $\Delta = \frac{L}{g(g-1)/2}$
- $C_B(n^*)$ = largest realized actor betweenness for the set of actors
- Betweenness Centralization: $C_B = \frac{2\sum_{i=1}^{g}\left[C_B(n^*) - C_B(n_i)\right]}{\left[(g-1)^2(g-2)\right]}$

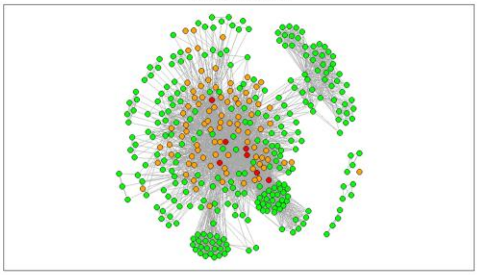
SHS, 1990 – 2011 (Excluding Big JAMA)

Author Role (PI, Co-PI, N)
SHS 1990-2011 (n=466)

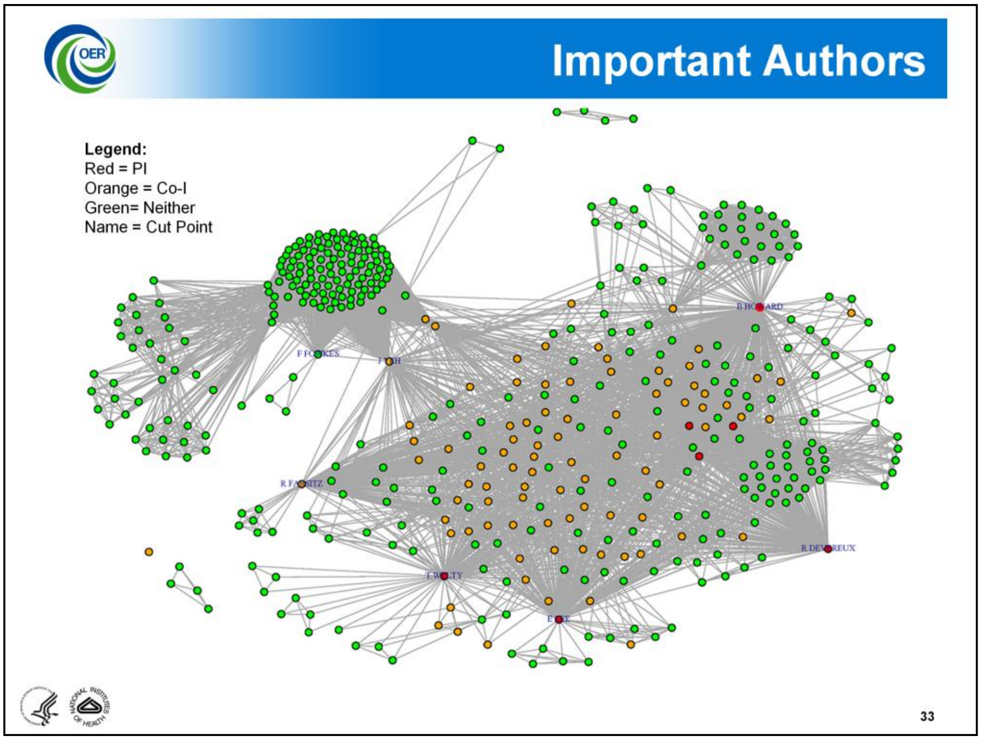


SHS:
Authors = 466
Density = 0.09
# Components = 4
Diameter = 4
Betweenness = 0.28

Author Status (PI, Co-P, N), excluding JAMA paper
SHS, 1990-2011 (n=372)

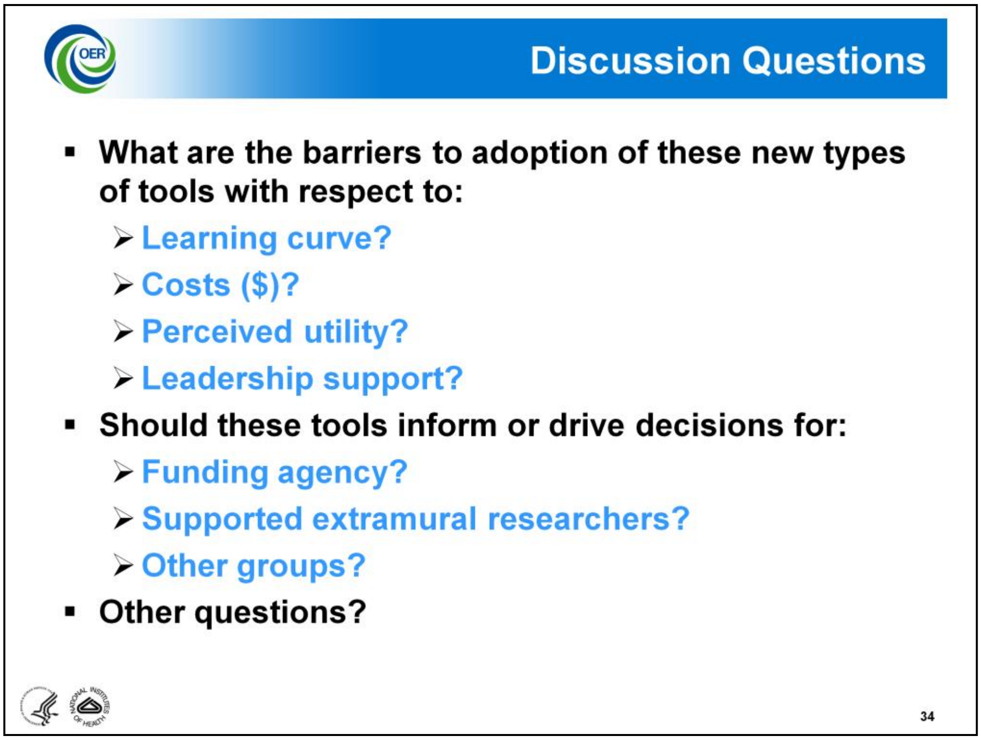


SHS:
Authors = 372
Density = 0.05
# Components = 5
Diameter = 5
Betweenness = 0.39



Important Authors

Legend:
Red = PI
Orange = Co-I
Green= Neither
Name = Cut Point



Discussion Questions

- What are the barriers to adoption of these new types of tools with respect to:
  - Learning curve?
  - Costs ($)?
  - Perceived utility?
  - Leadership support?
- Should these tools inform or drive decisions for:
  - Funding agency?
  - Supported extramural researchers?
  - Other groups?
- Other questions?